Metrics Brief

There are lies, there are damn lies, and then, there are statistics.

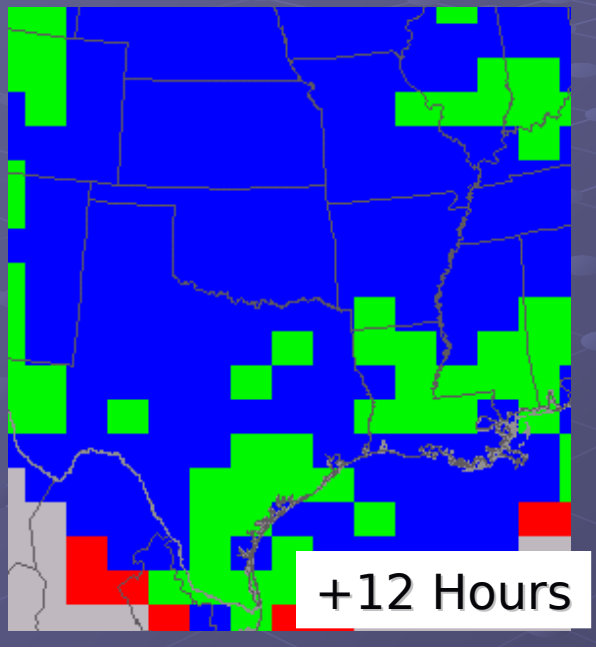
Mark Twain

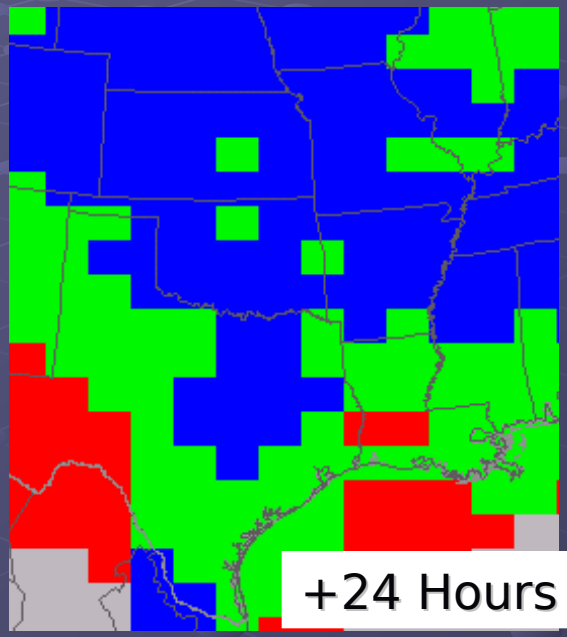
3 questions we want to answer:

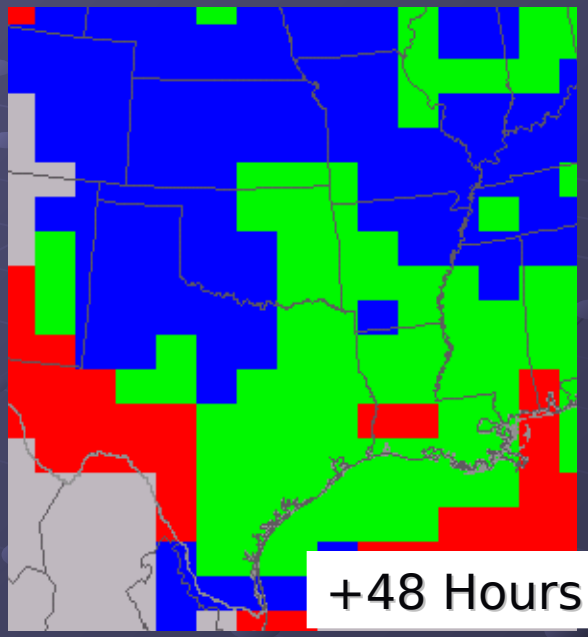
1. Did we collect a large enough sample to conclude our statistical measures are accurate?
   - Answer: *Chi-Square*
2. Do we forecast better than climatology, persistence, or pure randomness?
   - Answer: *Heidke Skill Score*
3. Do we tend to under-forecast or over-forecast phenomena?
   - Answer: *Bias Score*

Forecast Skill – HWD

Period: 01-31 Oct 02







Strong Skill (hss>.5)

Weak to Moderate Skill (0<hss<.5)

No Skill (hss< 0)

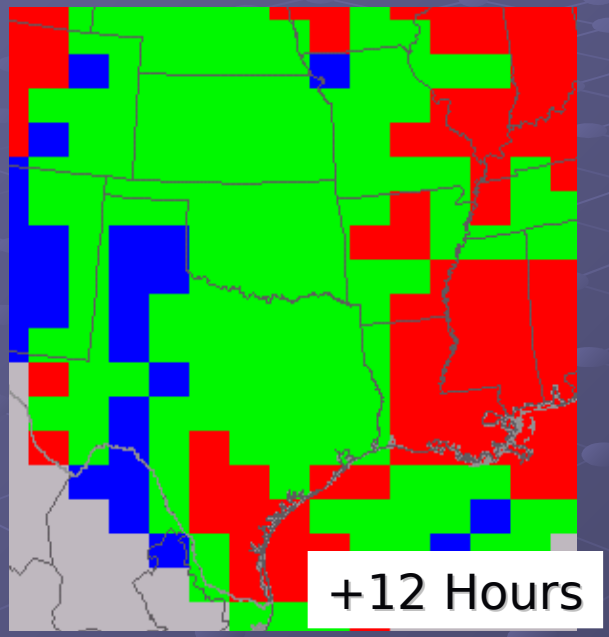
Insufficient Data

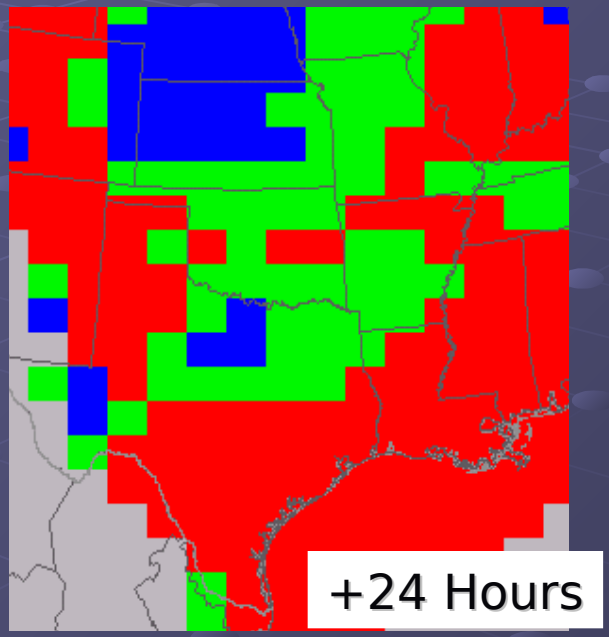
Negative HSS means the results are worse than totally random forecasts generated over the same monthly period.

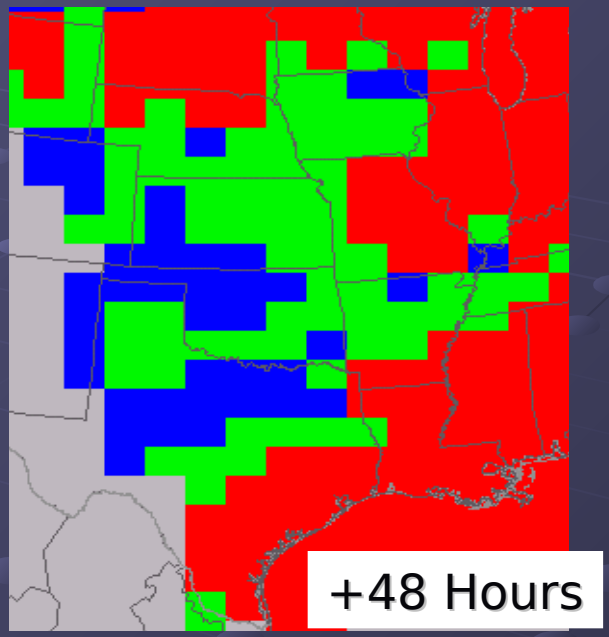
Climatology has a HSS of 10%

Bias – HWD

Period: 01-31 Oct 02







Over-Forecast (bias>1.2)

No Significant Bias (.8<bias<1.2)

Under-Forecast (bias< .8)

Insufficient Data

A bias score of 1 is perfect. Higher than 1 is positive bias and less than 1 is negative bias.

Chi-Square

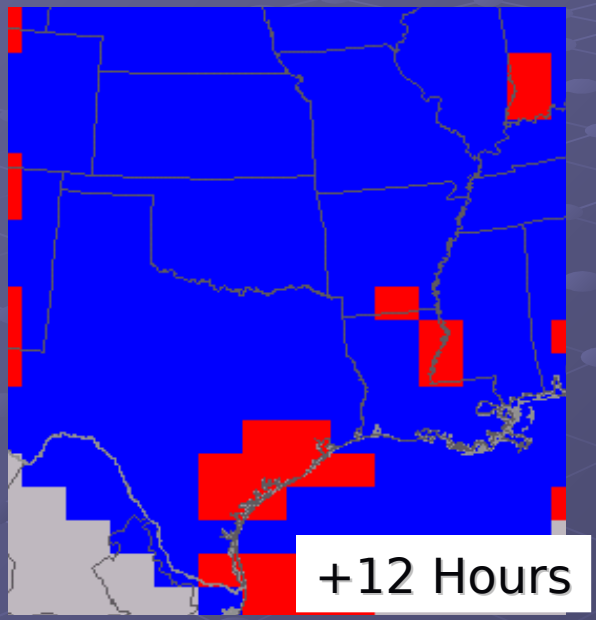
How we prevent our statistics from being worse than a “damn lie”.

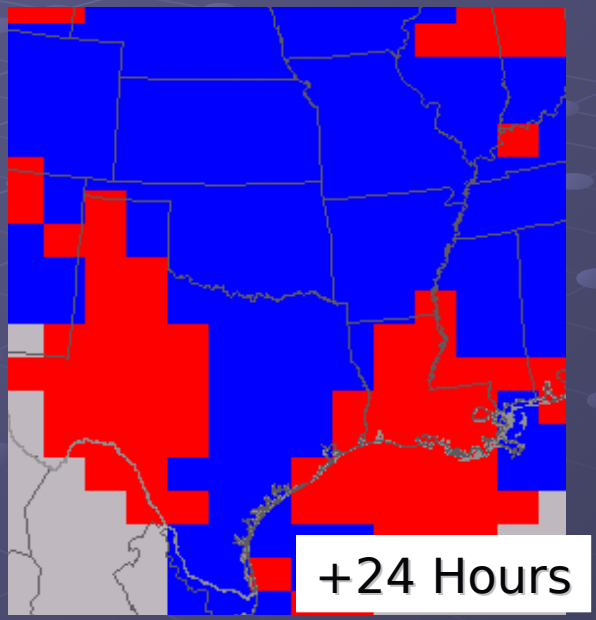
- Always the first step before concluding any results based on other categorical statistical measures.
- We have to make sure there is no large degree of randomness in our data that will skew our results.

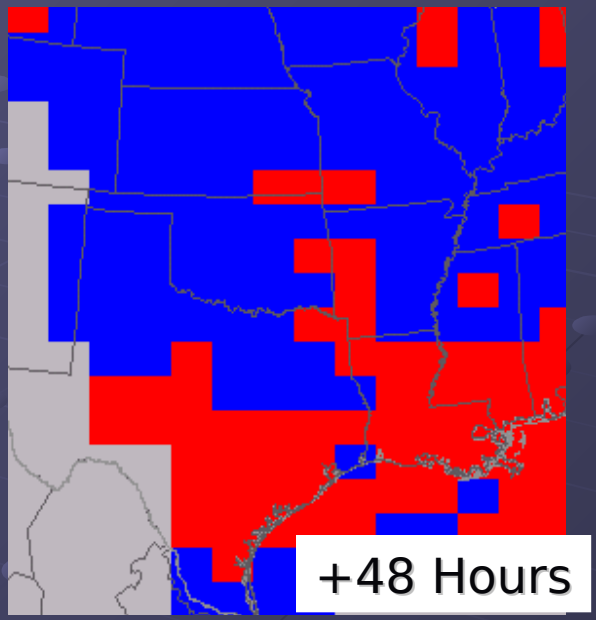
$$\chi^2 = \sum \frac{(\text{observed} - \text{expected})^2}{\text{expected}}$$

Chi-Square Distribution – HWD

Period: 01-31 Oct 02







Hypothesis: Are forecasts related in a meaningful way to corresponding observations?

- Yes (p<.05)
- No (p>.05)
- Insufficient Data

For each of these grids, the sample size is the same. However, the red grids have a certain degree of randomness such that we can't realistically conclude the results of other statistical measures (bias, skill, etc) are legitimate.

Chi-Square Fit Test Steps

- Start with the hypothesis: Our forecasts are related in a meaningful (non-random) way to their corresponding observations.
- Calculate Chi-Square value.
- Look up the corresponding p-value for this number in a statistics book (or use Excel).
- If the p-value ≥ .05, reject the hypothesis; if not, accept the hypothesis.

Statistical Dilemma

Sample A, size of 6:

	O-Yes	O-No
F-Yes	1	1
F-No	1	3

HSS: .25
Bias: 1

Chi-Square Statistic: .375
Chi-Square p-Value: .54

Sample B, size of 60:

	O-Yes	O-No
F-Yes	10	10
F-No	10	30

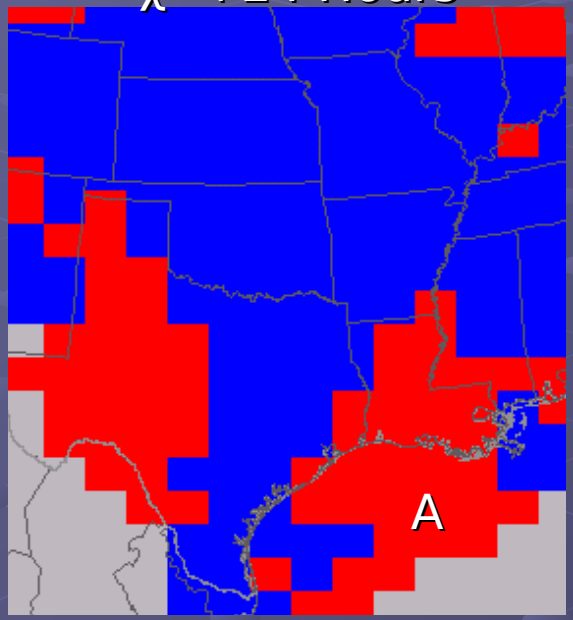
HSS: .25
Bias: 1

Chi-Square Statistic: 3.75
Chi-Square p-Value: .0528

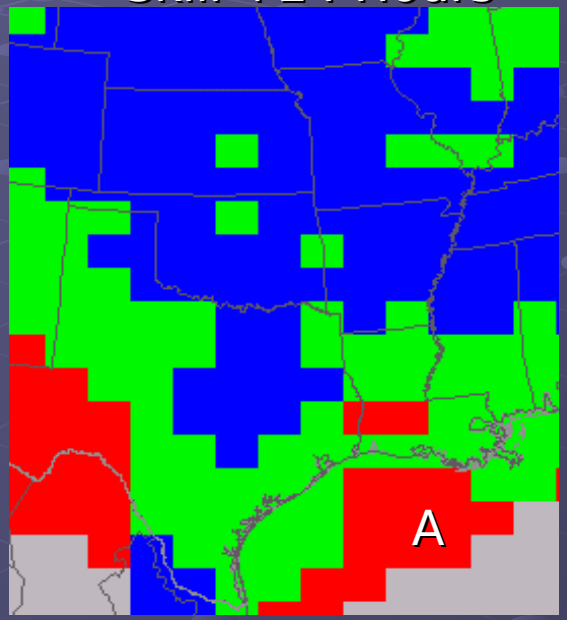
Although the skill and bias scores for both samples are exactly the same, we have to conclude the results from Sample A are bogus because the sample size of 6 is not large enough to make conclusive interpretation.

Rejecting Results

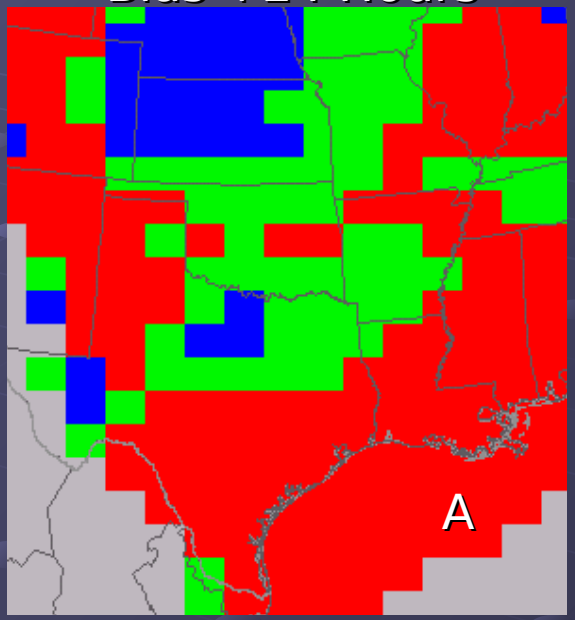
χ^2 +24 Hours



Skill +24 Hours



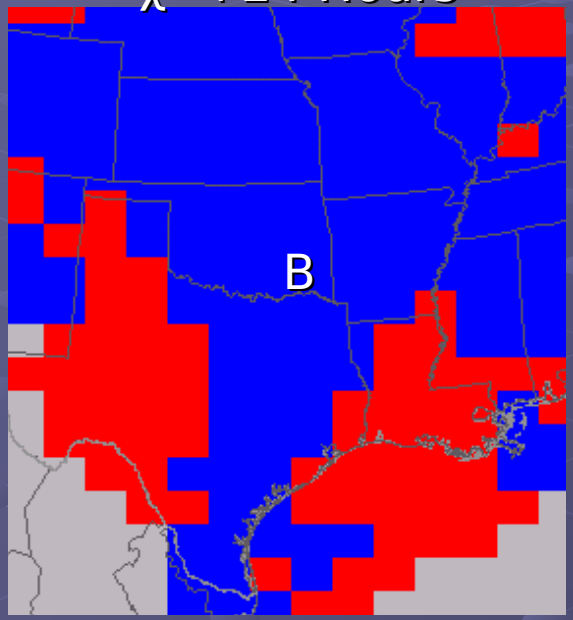
Bias +24 Hours



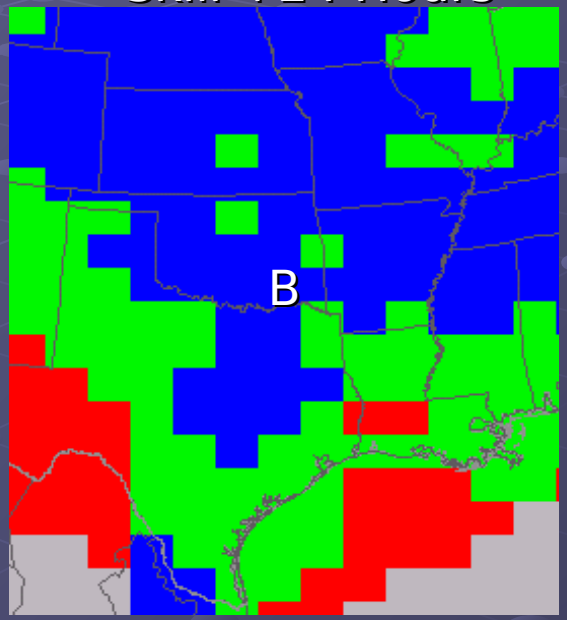
Point A – First impression is that we have low skill in forecasting IFR/MVFR in the Gulf with a under-forecasting bias being a contributing factor. However, the chi-square p-value is much greater than 5%. Our sample size is not large enough to wash out the randomness in our data, therefore, we don't have complete trust in our skill and bias numbers at point A.

Accepting Results

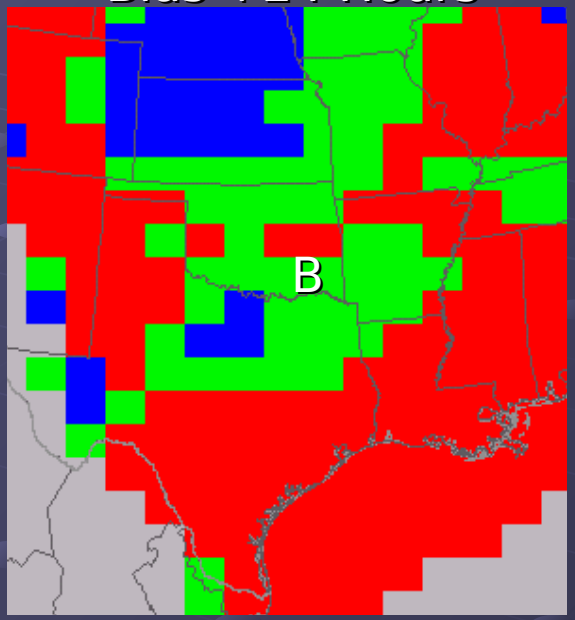
χ^2 +24 Hours



Skill +24 Hours



Bias +24 Hours



Point B – First impression is that we have high skill and insignificant bias in forecasting IFR/MVFR over SE.OK. The chi-square p-value, being less than 5%, supports these conclusions. Our sample size was large enough to wash out the randomness in our data, therefore, we have a high degree of trust in our skill and bias numbers at point B.